Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC workshop on *"Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges"*

Warsaw, Poland, 14-15 December 2015

In pursuit of patterns of economic behaviours using cluster and correspondence analysis[1]

Arkadiusz Florczak, Janusz Jabłonowski and Michał Kupc,
Narodowy Bank Polski (Poland)

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS or the central banks and other institutions represented at the meeting.

In pursuit of patterns of economic behaviours using cluster and correspondence analysis

The cluster and correspondence analysis with use of preselected features of the household budget survey with an automated documentation

Arkadiusz Florczak, Janusz Jabłonowski, Michał Kupc[1]

In pursuit for patterns of economic behaviours using cluster & correspondence analysis

Availability of large datasets of (often) sensitive data at the level of Statistics Department imposes the obligation to verify their quality and numerical consistency. However, such circumstances offer a chance to use the low level statistical tools based on purposefully written functions that search for unobservable patterns (clusters) that may not be apparent at higher levels of aggregation in publishable data. The versatile statistical tools created in R environment, relying in principle on e.g. cluster analysis and correspondence analysis, may serve in many fields as a link between micro and macro level analysis, with additional possibility to create the automatic documentation of the R code and results. The working example covers analysis of the saving behaviours based on household budget surveys.

Keywords: cluster analysis, correspondence analysis, household budget survey

JEL classification: C38, D14

[1] Authors are the employees of the Statistics Department in Narodowy Bank Polski.

Foreword

The statistics divisions settled in the domestic or international institutions are unique places of coincidence of special features: when looking from the input side there's nearly unconditional availability of often sensitive data for many sectors, with additional possibility to trigger the surveys, comparing them with the existing data. After a process of data modification and aggregation the output side (externally mainly websites) shows some loss in informative value (entropy), i.e. some units decide to publish mainly only the aggregates, while other disseminate more advanced statistical products. Simple transformation process usually covers verification of internal consistency, missing data, outliers and other comparable issues, then aggregation. Further levels of the data users receive the datasets theoretically free of numerical and logical errors and omissions. However, the pre-aggregated, user-friendly datasets may already trim unknown level of details theoretically interesting for analytics. Additionally, presentation of limited measures of the central tendency (e.g. arithmetic means) may often be misleading in non-normal distributions of the background data. The expansion of research can be promoted in statistical units by creation of statistical and econometric tools that not only verify raw source data for consistency and logical errors, but also searches for eventually hidden patterns in theoretically exploited topics. Since the authors are hired in the Statistics Department, there's a chance to create such toolbox to pre-aggregate and model easily accessible data. The toolbox formation started in early 2015 from, among others, cluster analysis and correspondence analysis, with imposed script formatting in open source R environment with additional aim for prompt automated documentation.

The structure of the paper: after current introduction there's a chapter on the methodology of the cluster and correspondence analysis extended by utility comments. Then the exemplary questions to be answered follow, and attempts of replies, including the box with a piece of automated documentation produced by the statistical tool. Next come the conclusions of the results, followed by the conclusions for the statistical tool and finally a literature review. The views expressed in the paper are not to be associated with those of the National Bank of Poland.

The idea of the cluster and correspondence analysis

The cluster analysis was introduced to the literature in 1939 by R. C. Tryon in "Cluster analysis". Due to large flexibility of application based on free-of-restrictions search for patterns of concentration of diverse features of the data, the clustering method became popular in many branches of science, e.g. biology, medicine, computer science, marketing or social sciences, including economics. In principle, the literature suggests 4 main methods of clustering, as follows:

1. hierarchical,
2. non-hierarchical,
3. graphical presentation,
4. hybrid.

Each method applies comparable stages of preparation, consisting of:

1. selecting objects and features that represent them,
2. transformation of the data,
3. choosing measures of distance,
4. selecting method of clustering,
5. selecting number of clusters,
6. evaluating of results of clustering,
7. interpretation and class profiling.

Strengths: A free of restrictions approach that seeks for the *natural* groups within complex sets of standardized data allows sometimes to find hidden structures and reduced number of explanatory variables, helpful when building e.g. a model that aims to simulate the reality. With longer time series of well-known data some new common features may occur in clusters. Multi-cluster structures can reveal e.g. different regional patterns.

Weaknesses: Firstly, the choice of number of classes can be a pejorative educated guess by a researcher. With large number of variables the choice of only one class may result in very condensed and large cloud of features, which is difficult for interpretation. Secondly, there is a difficulty in interpretation of the clusters even if they exist – since they might form an apparent concentration in terms of statistical structures. The comparison of several linking methods and distance measures shall help anyway. Thirdly, a result is prone to applied type of the distance measure.

The correspondence analysis, which notions dates back to early 20th century, was broadly popularized by Hill in 1974. It can be regarded as a special kind of a canonical correlation analysis between two categories of discrete data (Clausen, 1998). In principle, it's a graphical presentation of the relations between two or more sets of quantitative and qualitative data, usually in the form of (perceptual) maps. The strengths can be repeated after the cluster analysis, however, there are no limitations for the size of dataset that may consist of both variables and objects and the outliers do not affect the position of other objects on the map. The weaknesses may be extended by the suggestions of Greenacre (2009), e.g. time series analysis low frequency points are often situated in outlying positions in the map because of their unusual profiles.

The exemplary questions

Several reports on the Polish households suggest insufficient propensity to save (World Bank, Poland CEM, 2014), which, in line with the expected dropping replacement rates from the public pension system, may result in disappointing streams of income after the retirement. From the one hand, the households' saving motives seem theoretically quite well explained by the economics of consumer choice, but from the other hand, the observed financial savings' rate may suggest lack of strong saving motives for retirement. In pursuit for the savings' motives at the micro level the available household datasets for income and consumption may be handy – especially, if it occurred that some household features repeat more often when savings or excessive consumption are in focus. It could be also nice test

of statistical tool i.e. R based cluster and correspondence analysis, which could enrich knowledge at the macro level.

From the available datasets the household budget survey (HBS) was chosen, which is based on the representative method and covers the rotating sample of around 37 thousands households every year. The monthly rotation of households assumes that every month of the year a different group of households participates in the survey, and the ex-post stratification allows to generalize conclusions from samples into population every quarter, within a margin of an error. Each household participating in the survey keeps a special diary for a month, where registers expenditures, quantitative consumption and incomes[2].

The method of monthly rotation of the households' sample complicates the analyses of the time series when searching for patterns of economic behaviours in time. For modelling purposes short vectors of features are preferred or repeated observations the same units in time. The traditional approaches to modelling usually consider e.g. job experience and education, but is there a possibility that the vectors of the household features, which with higher propensity to consume / save, are longer and comprise dozens of features? Or in other words: is there a stable overtime set of HBS features that is common to households with high propensity to save (or consume)? What's the reduction of features for a model-saving household compared with the not-saving household, if any? How does the seasonality savings / excessive spending affect a composition of clusters? Do indebted households share common features and behaviours with those free of mortgage? Does mortgage currency affect clusters?

An attempt to reply

Since the main aim was to verify if there are overtime relatively stable set of features of households that create the highest savings (or excessively consume), the deciles of the annual cash result of employees' households[3] served for a basis for the initial stage grouping. The annual cash result is here calculated as a disposable income minus expenses (according to HBS methodology, the expenses category excludes the accumulation, e.g. real estate purchases, but include consumption of fixed capital, i.e. renovation) repeated in the following years and quarters between 2005 and 2014. Other words, the 1st decile covers the most indebting households, while the 10th comprises the most saving ones. The saving households' group grows with years, however, a number of households in each decile group is quite comparable when considering the generalization of the HBS sample on the entire population with weights.

During a stage of data transformation the quantitative data were transformed into qualitative data that reflected e.g. deciles of income, expenditures and their difference (referred here as voluntary savings) while e.g. the age vector grouped into labelled 5-year cohorts. Such method sacrifices the clarity of the results (large

---

[2] Source: http://stat.gov.pl/obszary-tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/budzety-gospodarstw-domowych-w-2014-r-,9,9.html.

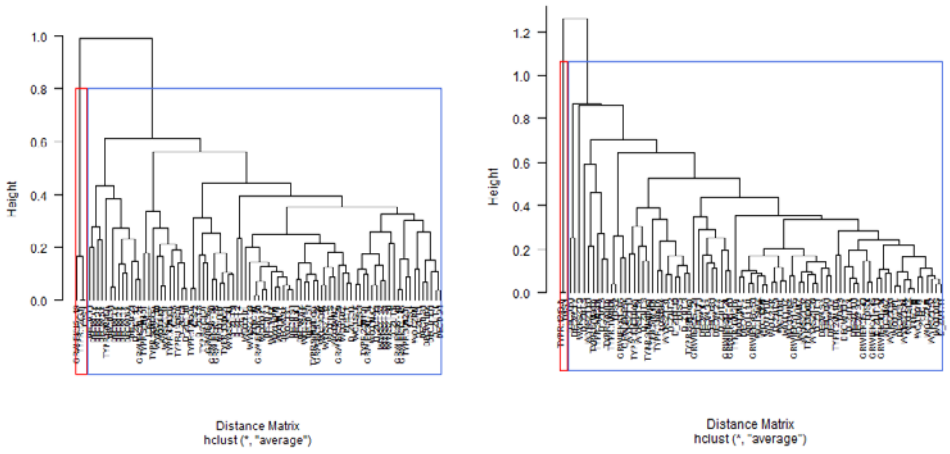
[3] Such distinction seems vital from the point of view of households composed of e.g. an employee and an entrepreneur, where incomes would be affected by a turnover from economic activity, and part of private consumption would be deformed by costs of entrepreneurship.

and dense clouds) but should be more profitable in terms of precision. Due lack of acceptable clarity of the results, the cluster analysis was followed by the correspondence analysis, where the data were transformed into artificial data in process of correspondence analysis.

When choosing the distance measures during the cluster analysis, for the binary data the Jaccard measures of dissimilarity, while for the non-binary data (qualitative), the Sneath coefficient, the Gower dissimilarity measure and Sokal & Michener and GDM2 measures were applied. In the following correspondence analysis, due to lack of the literature support for the non-hierarchical methods, only the hierarchical methods were considered with the Euclidean distance measure.

For the optimal choice of the hierarchical method of the cluster analysis followed by the correspondence analysis, the results were evaluated in terms of their deformation after converting distance matrix into cophenetic matrix. The verification relied on the two indicators: the cophenetic correlation coefficient and the sum of square deviations. In the analysed case both indicators suggested the average linkage method as optimal. The visual interpretation of the dendrogram, when evaluating the optimal number of groups, may be tricky, therefore, a further support can be found in 4 relative criteria: silhouette index (SIL), Caliński&Harabasz index (G1), Krzanowski&Lai index (KL) and Davies-Bouldin index (DB). The aim of the profile indices is to evaluate a theoretical appropriateness of selected number of clusters.

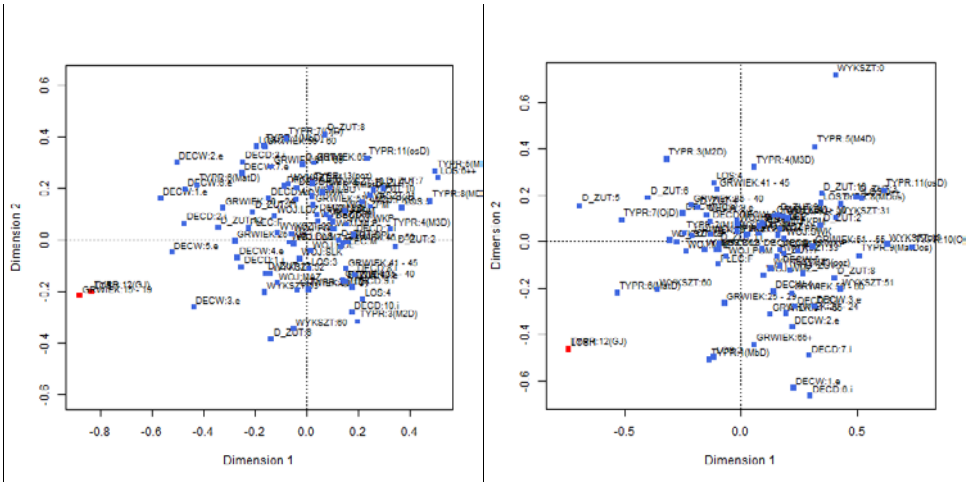
Picture 1 and 2 visualizes the cluster analysis via dendrograms for the employees households' features clustered according to the absolute value of deciles of their annual cash result: 1st (left) and 10th decile (right), with average linkage method for 2014 data.



Pictures 1 and 2, despite being blurred in attached version, share common main interpretation: the blue box covers nearly all 99 analyzed features of the households for 1st and 10th decile of annual cash result in 2014, so any concentration of features for model saving households, based on the HBS, does not occur in this version cluster analysis. The red field in dendrograms covers the outliers, expectedly consisting of youngsters, elders and low income.

Picture 3 and 4 below represent the same features with use of correspondence analysis, presented in a form of the perceptual maps for the same group of

employees' households, grouped in deciles of their annual cash result: 1st (left) and 10th decile (right), average linkage method for 2014 data.



The perceptual maps (result of correspondence analysis) confirm a very modest differentiation of features between the most saving and excessively spending employees' households. An alternative combination of available linking methods, types of distances and indices gives comparable results and additionally is not optimum from the point of view of applied indicators.

Clustering for households with mortgages in CHF extended for quarterly data

This part repeats the above described clustering procedure for the narrower part of the HBS, namely, households with mortgage loans in Swiss francs (CHF). A motivation for this particular trimming of input stems from the expectation that features of the indebted household may vary between annual savers and deficit makers due to higher insolvency risk. For instance the CHF indebted households with higher education level of head of family and low decile of disposable income would be rather risk averse and stabilize their budget or create a buffer stock for unpleasant currency shocks, rather than excessively spending and putting their household at risk. After the first round of calculations a question occurred if some more information can be achieved from the trimmed set of quarterly HBS data. A rationale of this additional exercise has risen from the possible effect of the monthly rotation of the households in the HBS sample, which in annual data could create noise in clusters due to e.g. overlapping seasonal effects in income fluctuations and consumption decisions.

Incurring mortgage loans in Swiss francs (CHF) supposed to be popular and broadly available few years back in Poland. Three major shocks in CHF/PLN from 2009, 2011 and 2015 coincide the households' growing insolvency (macro-prudential statistics) revealed that a part of them was not sufficiently prepared for the exchange rate shocks, despite dropping interest rates. Starting from 2012 the HBS covers also the currency for mortgage and loans, therefore, up to 2014 complete 12 sets of quarterly panel data are available. The literature and media repeats generally well known features of the households with mortgages in CHF, i.e. 2+2 composition, living in towns over 500 thousands inhabitants, occupying an apartment below

75m^2, in a building raised between 1996 and 2006. Interestingly, despite short coverage period (3 years), and modest representation (2,4% of all households), a significant linear decrease in number of households with CHF mortgage can be observed: by 20% between 2012 and 2014, yet hardly explained, while declared mortgage repayment period averaged to 24 years.

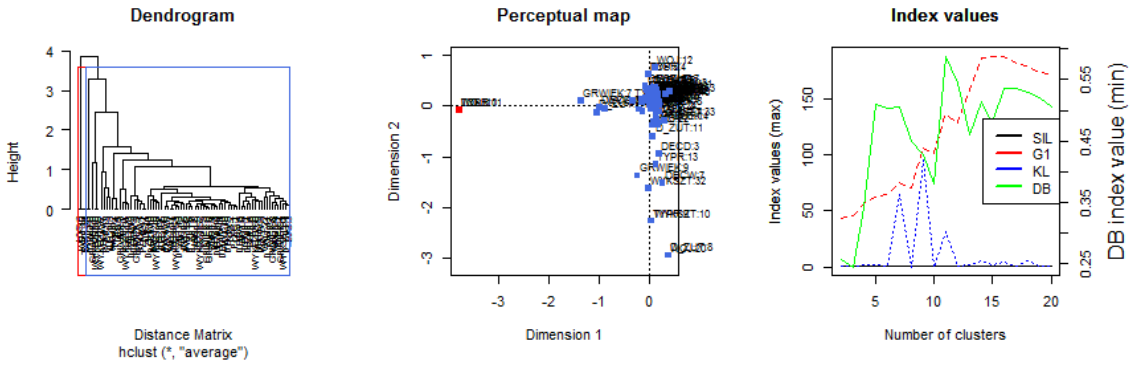
Additionally, an example of automated documentation is inserted in box below, based on Markdown: a lightweight markup language triggered from the level of R compiler, i.e. R Studio that helps to produce fancy and even interactive e.g. HTML, PDF, Word documents and presentations.

The box shows a piece of exemplary automated output in a form of explanatory text, charts and desired piece of R code specified on grey stripes called "chunks":

1

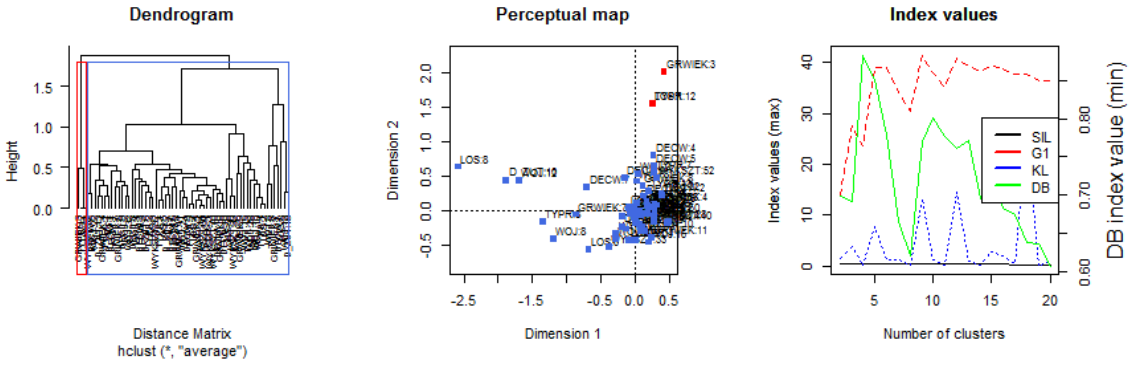
Example of automated documentation in .doc file
[…]
The dendrogram (left) and the perception map (center) profile index (right) with differently colored clusters obtained by average linking method. The chart on the right hand side shows the changing profile indices with growing number of clusters. There are possible profits of playing with the number of clusters, e.g. if for each cluster in HBS analysis different regions or income deciles are falling. Growing number of clusters (limited here to 20) also gives a possibility of the reduction of the explanatory variables in each cluster that can be easier used in econometric modelling, if it seems to make sense, of course.
Sequences of pictures below show the decile for the households with the highest annual deficit in 2014:

```
a14.1 <- an_kores_skup(dane = dane,rok = "2014", dec = "1", odl="euclidean", metoda_skupien="average",max_nc=20)
```



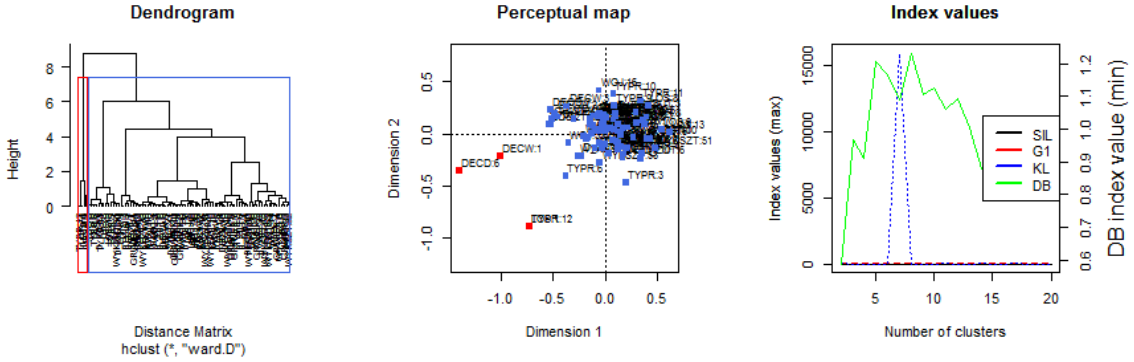
Below the same set for decile with the highest savings:

```
a14.10 <- an_kores_skup(dane = dane,rok = "2014", dec = "10", odl="euclidean", metoda_skupien="average",max_nc=20)
```

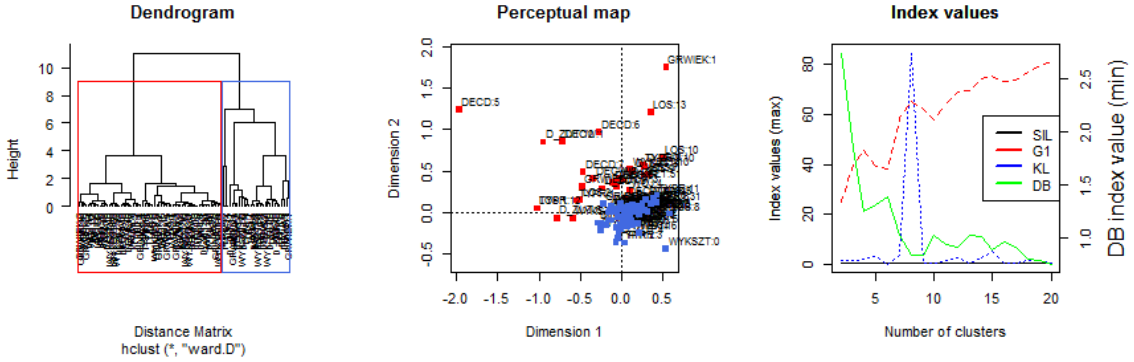


The HBS is stratified in quarterly periods, which allows to reweight the sample for the entire population on quarterly basis too. While the annual data cover all 37 thousands households, so the strong concentrations can be spotted, but at the same time strong seasonal impulses form the income can overlap. There is then a possibility that the cluster and correspondence analysis reflect some seasonal patterns. The following sequence of graphs shows the quarterly sequence of clusters for the most saving households bearing mortgage in 2014 (10th decile of quarterly cash result):

1st quarter:

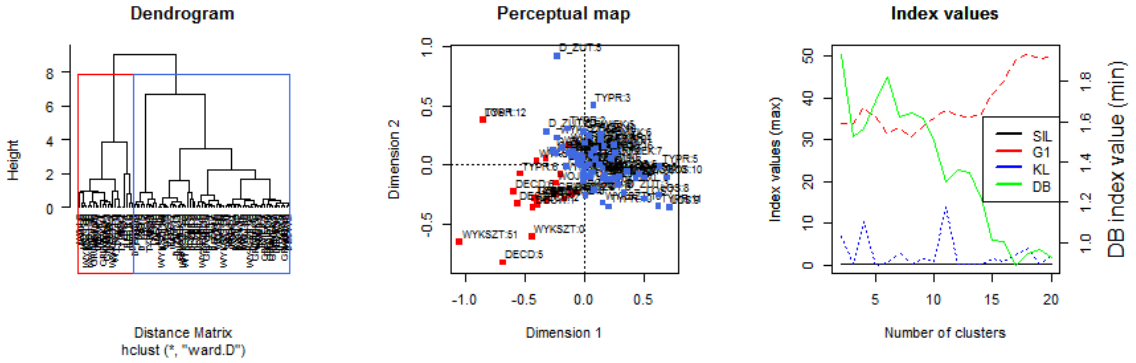


2nd quarter:

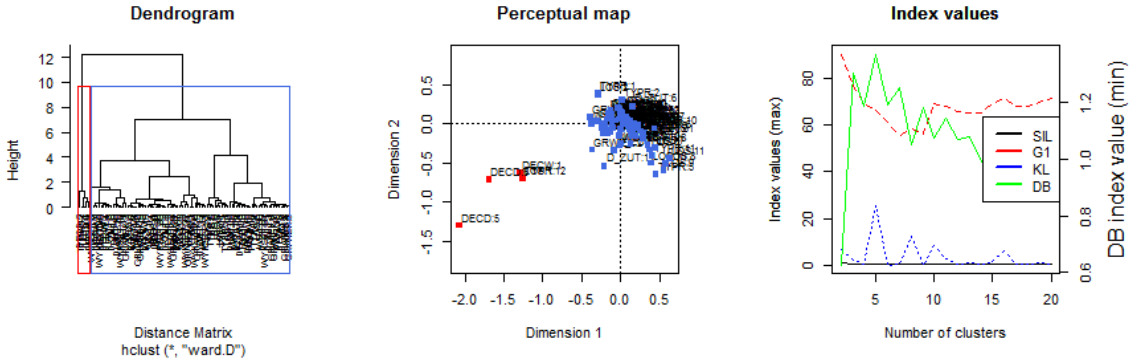


1

3rd quarter:



4th quarter:



Conclusions of the results

The initial aim to find the small sets of (optimally) interpretable features for the most saving and most indebting households was not achieved. The results of the HBS based cluster and correspondence analysis applied on 99 normalized features of employees' households show no difference between two marginal types of households in terms of their annual cash results: in fact single large cluster occurred. Additional analysis for the narrower group of households with mortgage in CHF also shows no significant differences between the most indebting and most saving households. Further trimming of the input dataset into quarters increased the set of dendrograms, perceptual maps, index charts and referred conclusions beyond initially assumed range of this paper. Some seasonal patterns occurred, while the outliers remained in clusters despite further narrowing of the input dataset. The differences in clusters between mortgage indebted and not indebted households finally occurred in quarterly data too, however, not interpretable for authors at this stage of analysis.

In further pursuit for stable behaviour patterns, the results may suggest to extend the number of features rather than searching among the existing ones. If households' consumption reflects somehow a seek for satisfaction of needs, then maybe differentiation in terms of households' subjective evaluation of living

conditions (satisfaction of needs in Maslow hierarchy), also reported in the HBS since 2005, covers a quality of inhabited apartment or house. Some interesting initial results, not considered in this paper, occur also for a differentiation of period of a life cycle, i.e. for pensioners' households.

Initial evaluation of applicability of the statistical tool

With regard to the described part of the data mining statistical tool, i.e. cluster and correspondence analysis, their initial preparation from scratch took around 400 hours of literature review, coding in R and ordered documenting of the knowledge and experience database. Further extension of the input can be nearly immediate, however, depends on the nature of analysed feature. The tool suits for different input data, e.g. other surveys, administrative data and matched datasets, however, the process of normalisation, decision tree for choosing distance measures, selecting clustering methods and evaluation of its appropriateness still requires further automation. It was initially tested with the HBS set – the largest and most complex on the household data. The quantitative data generated in clustering procedure may be very helpful in difficult imputations between e.g. administrative and survey datasets, e.g. if obvious common features or distance measures are insufficient. Perhaps, it may serve as a last resort help in reduction of number of potential explanatory variables in modelling, where e.g. number of parameters is close to or exceeds a number of degrees of freedom, or where e.g. various Lasso's fail due to type or size of data.

The function of automated documentation suits many applications at all stages of the computation results. There are very promising examples in initial verification of the raw dataset quality using also very efficient algorithms based on 'Rcpp' package. The Markdown 'shiny' option offers interactive websites, which can be handy especially if the results exceed basic comprehension and can't be further reduced without sacrificing of the essential findings. In this paper the automated Word files allowed for an immediate presentation of uncertain results at 40 pages of charts in trusted order and coherent layout without problems known from copy & paste procedure.

Literature

1. Brzezińska J, (2011), Analiza korespondencji, [w:] Gatnar E., Walesiak M. (red.), Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R, Wydawnictwo C. H. Beck, Warszawa.
2. Buko A. (1991), Analiza skupień w badaniach wczesnośredniowiecznych surowców garncarskich: przykład ceramiki sandomierskiej, Polska Akademia Nauk. Instytut Historii Kultury Materialnej, Kraków.
3. Deaton A. (1997), The Analysis of Household Surveys, A Microeconometric Approach to Development Policy, Johns Hopkins University Press for the World Bank, Baltimore.
4. Dębkowska K. (2010), Metody statystyczne w segmentacji rynku, „Ekonomia i Zarządzanie, nr 4.
5. Gatnar E., Walesiak M. (2011), Analiza danych jakościowych i symbolicznych z wykorzy-staniem programu R , Wydawnictwo C. H. Beck, Warszawa.
6. Gospodarstwa domowe i rodziny, Narodowy Spis Powszechny Ludności i Mieszkań. Powszechny Spis Rolny (2003), GUS, Warszawa.
7. Gower J. C. (1971), A General Coefficient of Similarity and Some of Its Properties, Biomet-rics, vol. 27, nr 4.
8. Hartigan J. A., Wong M. A. (1979), K-Means Clustering Algorithm, „Journal of the Royal Statistical Society. Series C (Applied Statistics)", vol. 28, nr 1.
9. Huang Z. (1997), A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. in KDD: Techniques and Applications (H. Lu, H. Motoda and H. Luu, Eds.), World Scientific, Singapore.
10. Kisielińska J. (2009), Bezwzorcowa klasyfikacja obiektów w ekonomice rolnictwa, „Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie" 2009, Tom 8 (XXIII), Wydawnictwo SGGW, Warszawa.
11. Kolasa-Więcek A., Tukiendorf M. (2012), Zastosowanie metod grupowania aglomeracyjnego w segmentacji państw Unii Europejskiej, „Journal of Research and Applications in Agricultural Engineering", vol. 57 (1).
12. Laskowski W. 2005, Studium realizacji potrzeb żywnościowych ludności Polski na tle wielo-wymiarowych klasyfikacji i analiz gospodarstw domowych, Wydawnictwo SGGW, Warszawa.
13. Marbac M., Biernacki C., Vandewalle V. (2014), Model-based clustering for conditionally correlated categorical data, Rapport de recherche INRIA RR-8232.
14. Marbac M., Biernacki C., Vandewalle V. (2014), Finite mixture model of conditional de-pendencies modes to cluster categorical data, Reprint.
15. Migdał-Najman K. (2011), Ocena jakości wyników grupowania – przegląd bibliografii, „Przegląd Statystyczny", Zeszyt 3-4 (R. LVIII).
16. Migdał-Najman K., Najman K. (2013), Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej, „Zarządzanie i Finanse", nr 3.

17. Panek T. (2015), Analiza porównawcza subiektywnego dobrostanu w Europie, „Wiadomości Statystyczne GUS", nr 2 (645).

18. Panek T., Zwierzchowski J. (2013), Statystyczne metody wielowymiarowej analizy porów-nawczej. Teoria i zastosowania, Oficyna Wydawnicza SGH w Warszawie, Warszawa.

19. Piekut M. 2008, Polskie gospodarstwo domowe – dochody, wydatki i wyposażenie w dobra trwałego użytkowania, Wydawnictwo SGGW, Warszawa.

20. Podogrodzka M. (2011), Analiza zjawisk społeczno-ekonomicznych z zastosowaniem metod taksonomicznych, „Wiadomości Statystyczne GUS", nr 11 (606).

21. Salamaga M. (2009), Analiza zróżnicowania struktury wydatków gospodarstw domowych, „Wiadomości Statystyczne GUS", nr 5 (576).

22. Struyf A., Hubert M., Rousseeuw P. J., Clustering in an Object-Oriented Environment, Department of Mathematics and Computer Science, U. I. A., Antwerp.

23. Studnicki M., Mądry W., Śmiałowski R. (2009), Porównanie efektywności metod staty-stycznych tworzenia kolekcji podstawowej na przykładzie pszenicy jarej, „Biuletyn Instytutu Hodowli i Aklimatyzacji Roślin", nr 252.

24. Tian B., Kulikowski C. A., Leiguang G., Bin Yamg, Lan Huang, Chunguang Zhou (2012), A Global k-modes Algorithm for Clustering Categorical Data, „Chinese Journal of Electronics", vol. 21, nr 3.

25. Walesiak M. (2012), Klasyfikacja spektralna a skale pomiaru zmiennych, „Przegląd Statystyczny", Zeszyt 1.

26. Walesiak M. (2003), Miara odległości obiektów opisanych zmiennymi mierzonymi na różnych skalach pomiaru, [w:] Zastosowania statystyki i matematyki
w ekonomii, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1006, Wrocław.

27. Walesiak M. (2002), Pomiar podobieństwa obiektów w świetle skal pomiaru i wag zmiennych, Ekonometria 10, Prace Naukowe Akademii Ekonomicznej we Wrocławie nr 950, Wrocław.

28. Walesiak M. (2004), Problemy decyzyjne w procesie klasyfikacji zbioru obiektów, Eko-nometria 13, Prace Naukowe nr 1010 Akademii Ekonomicznej we Wrocławiu, Wrocław.

29. Walesiak M. (2005), Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji, Taksonomia 12. Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe nr 1076 Akademii Ekonomicznej we Wrocławiu, Wrocław.

30. Walesiak M. (2008), Procedura analizy skupień z wykorzystaniem programu komputerowego clusterSim i środowiska R, Taksonomia 15. Klasyfikacja i analiza danych – teoria i zasto-sowania, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7 (1207), Wrocław.

31. Walesiak M. (2014), Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej, „Przegląd Statystyczny", Zeszyt 4 (R. LXI).

32. Walesiak M. (2011), Uogólniona miara odległości GDM w statystycznej analizie wielowy-miarowej z wykorzystaniem programu R, Wydawnictwo Uniwersytetu Ekonomicznego we Wro-cławiu, Wrocław.

33. Walesiak M., Dudek A. (2009), Ocena wybranych procedur analizy skupień dla danych po-rządkowych, [w:] Taksonomia 16. Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Wrocław.

34. Walesiak M., Dudek A. (2006), Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu, [w:] Prace Katedry Ekonometrii i staty-styki nr 17, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 450, Szczecin.

35. Walesiak M., Gatnar E. (2004), Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

36. Ward J. H. (1963), Hierarchical grouping to optimize an objective function, „Journal of the American Statistical Association", vol. 58.

37. Warzecha K. (2009), Poziom życia ludności Polski i pozostałych krajów Unii Europejskie – analiza taksonomiczna, [w:] Pangsa-Kania S., Szczodrowski G. (red.), Gospodarska polska po 20 latach transformacji: osiągnięcia, problemy i wyzwania, Wydawnictwo Instytut Wiedzy i Innowacji, Warszawa.

38. World Bank (2014), Poland - Country economic memorandum: Saving for growth and prosperous aging, Washington, D.C.

39. Wołodźko T., Kokoszka A. (2014), Próba klasyfikacji osób podejmujących zachowania samobójcze – przegląd badań z zastosowaniem analizy skupień, „Psychiatria Polska, nr 4 (48).

40. Zhang B., Sargur N. S., Properties of Binary Vector Dissimilarity Measures, CEDAR, Computer Science and Engineering Department State University of New York at Buf-falo, Buffalo, NY 14428.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC workshop on *"Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges"*

Warsaw, Poland, 14-15 December 2015

In pursuit of patterns of economic behaviours using cluster and correspondence analysis[1]

Arkadiusz Florczak, Janusz Jabłonowski and Michał Kupc,
Narodowy Bank Polski (Poland)

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS or the central banks and other institutions represented at the meeting.

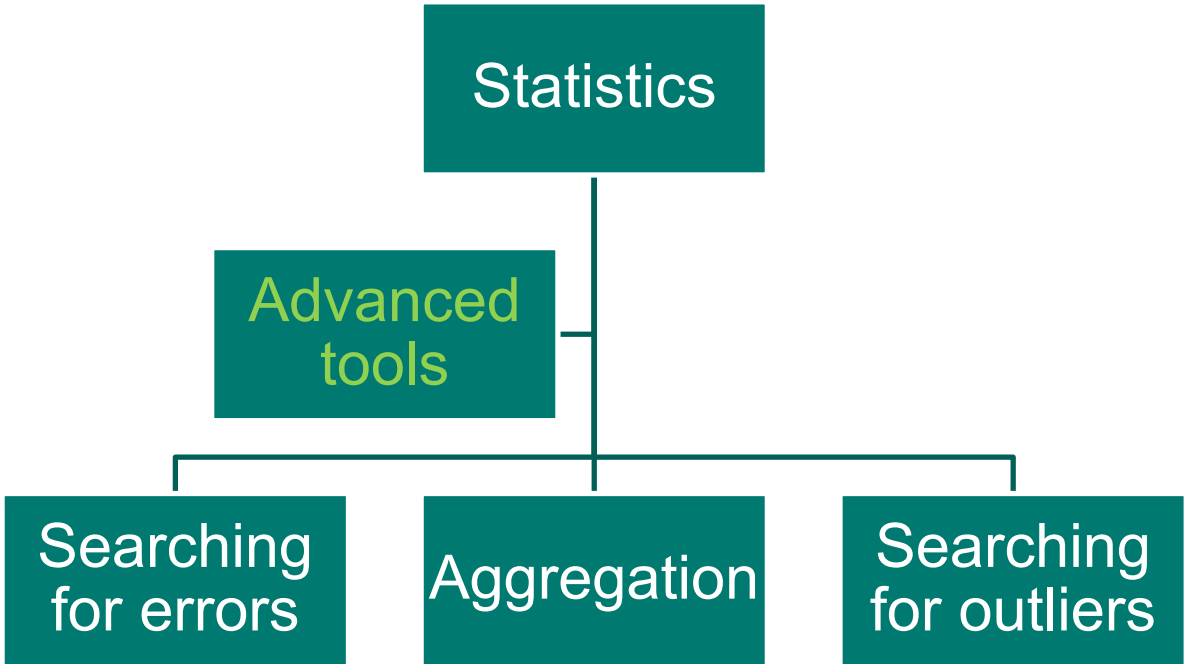
NBP Narodowy Bank Polski

Arkadiusz Florczak, Janusz Jabłonowski, Michał Kupc

In pursuit of patterns of economic behaviours using cluster and correspondence analysis

Warsaw / 14 December 2015

Statistics divisions



Advanced analysis

- Searching for new patterns in theoretially exploited datasets.
- Promoting research (macro level) based on more detailed data (micro level).

- Purposful building of statistical and econometric toolbox, started in early 2015.
- Universality, free of restrictions, matching datasets, imputing missing data
- Created from scratch in R environment (open source, free of charge, ranks high in IEEE), automated reporting via Markdown.
- Demand building: firstly internally, if tests go well, externally.

Exemplary tool: cluster analysis

- Based on free-of-restrictions search for patterns of concentration of diverse features of the data.
- Popular in many branches of science, e.g. biology, medicine, computer science, marketing or social sciences, including economics.

- **4 main methods of clustering:**
  - hierarchical,
  - non-hierarchical,
  - graphical presentation,
  - hybrid.

- **Procedure:**
  - selecting objects and features that represent them,
  - data transformation,
  - measures of distance,
  - method of clustering,
  - number of clusters,
  - evaluation of clustering,
  - interpretation and class profiling.

Exemplary tool: correspondence analysis

- Special kind of a canonical correlation analysis between two or more categories of discrete data.
- Graphical presentation of the relations between two or more sets of quantitative and qualitative data, usually in the form of (perceptual) maps.

- **Strengths:**
  - allows sometimes to find hidden structures in theoretically exploited data,
  - there are *no limitations* for the size of dataset,
  - dataset may consist of both variables and objects.
- **Weaknesses:**
  - choice of number of groups can be a pejorative educated guess,
  - difficulty in interpretation of the clusters.

The test on household budget survey (HBS)

- Several reports on the Polish households suggest insufficient propensity to save (World Bank, Poland CEM, 2014).
- The HBS as input data: rotating sample of around 37 thousands households every year, 2005-2014, around 200 features (if large set of detailed consumption categories excluded).
- 99 features included in the exercise.

Questions:

- Is there a possibility that the vectors of the household features, which with higher propensity to consume / save, are longer and comprise dozens of features?
- How does the seasonality savings / excessive spending affect a composition of clusters?
- Do indebted households share common features and behaviours with those free of mortgage?
- Does mortgage currency affect clusters?

Input preparation

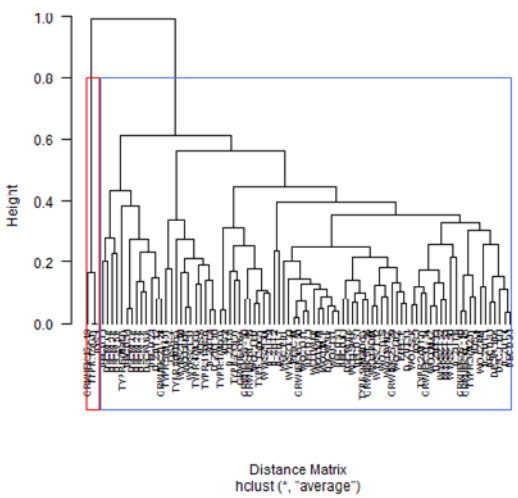
- Initial HBS grouping: the decile groups of the annual cash result of employees' households: 1st decile: the highest deficit, 10th decile: the highest savings.
- Cash result is here calculated as a disposable income minus expenses.
- The expenses category excludes the accumulation, e.g. real estate purchases, but include consumption of fixed capital, i.e. renovation.
- The saving households' group grows with years, however, a number of households in each decile group is quite comparable.

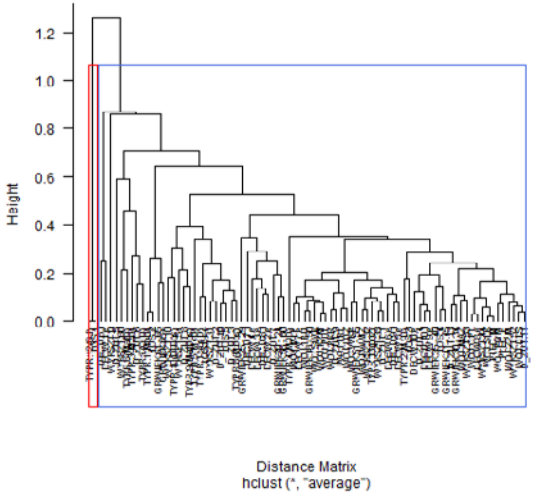
Clustering procedure

- During a stage of data transformation the quantitative data were transformed into qualitative data that reflected e.g. deciles of income, expenditures and their difference (referred here as voluntary savings).
- Distance measures: the Jaccard measures of dissimilarity for binary data, while for the non-binary data (qualitative), the Sneath coefficient, the Gower dissimilarity measure and Sokal & Michener and GDM2.
- Clustering: hierarchical and non-hierarchical methods.
- Correspondence analysis: hierarchical methods.
- For the visual interpretation of the perception maps 4 indices were used: profile index (SIL), Caliński&Harabasz index (G1), Krzanowski&Lai index (KL) and Davies-Bouldin (DB).

Cluster analysis results for annual cash result

- Picture 1 and 2 visualizes the cluster analysis via dendrograms for the employees households' features clustered according to the absolute value of deciles of their annual cash result: 1st (left) and 10th decile (right), with average linkage method for 2014 data.



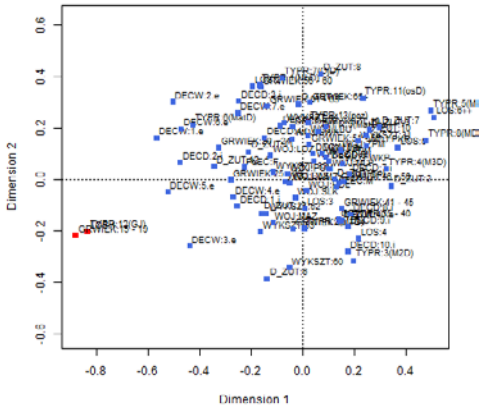


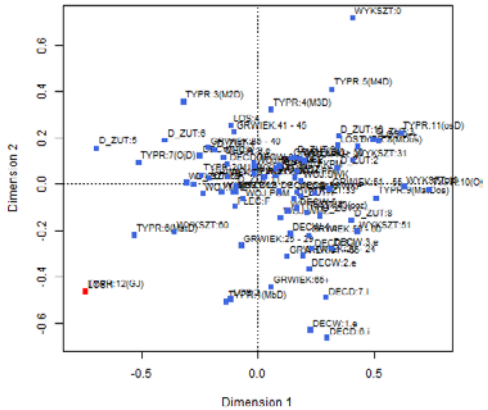
No differences!

Source: Own calculations.

Correspondence analysis results for annual cash result

- Picture 3 and 4 visualizes the correspondence analysis for the employees households' features clustered according to the absolute value of deciles of their annual cash result: 1st (left) and 10th decile (right), with average linkage method for 2014 data, euclidean distance.



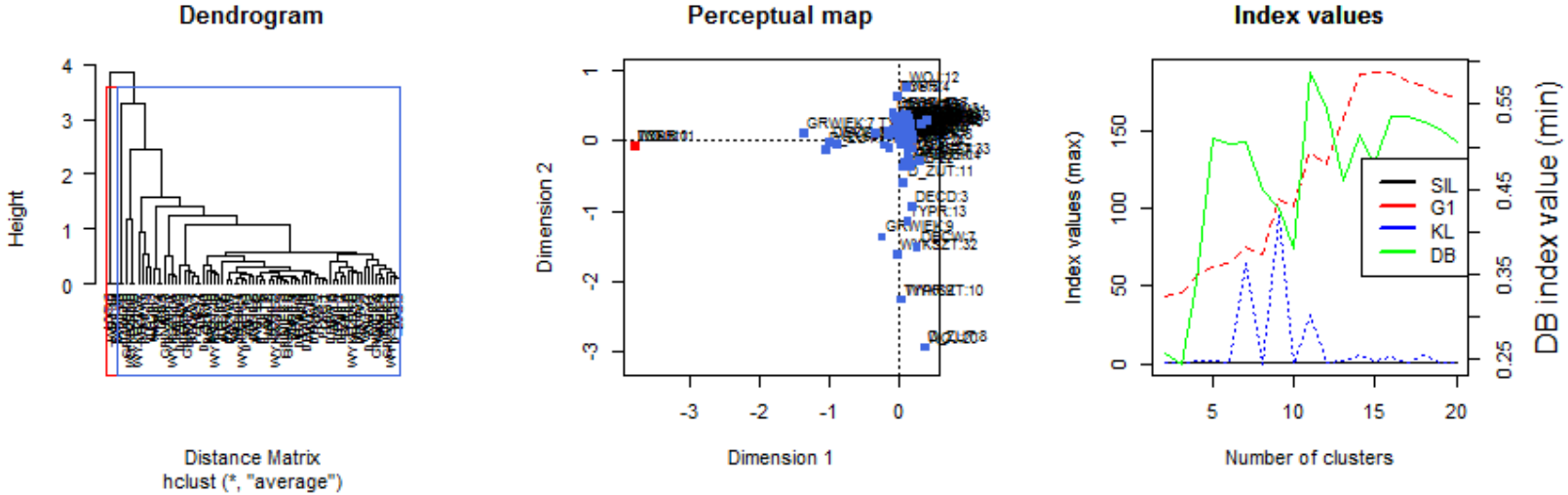


No differences as well!

Source: Own calculations.

Clustering for households with mortgages in CHF

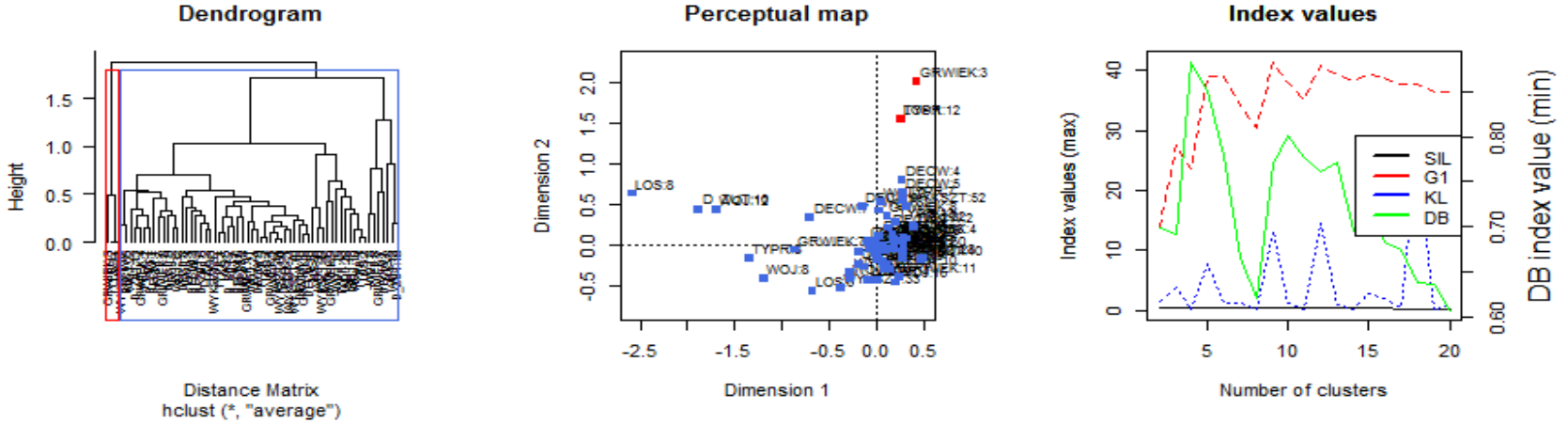
- The dendrogram (left) and the perception map (center) profile index (right) with differently colored clusters obtained by average linking method. **Deficit** incurring households with mortgage in CHF in 2014.



Source: Own calculations.

Clustering for households with mortgages in CHF

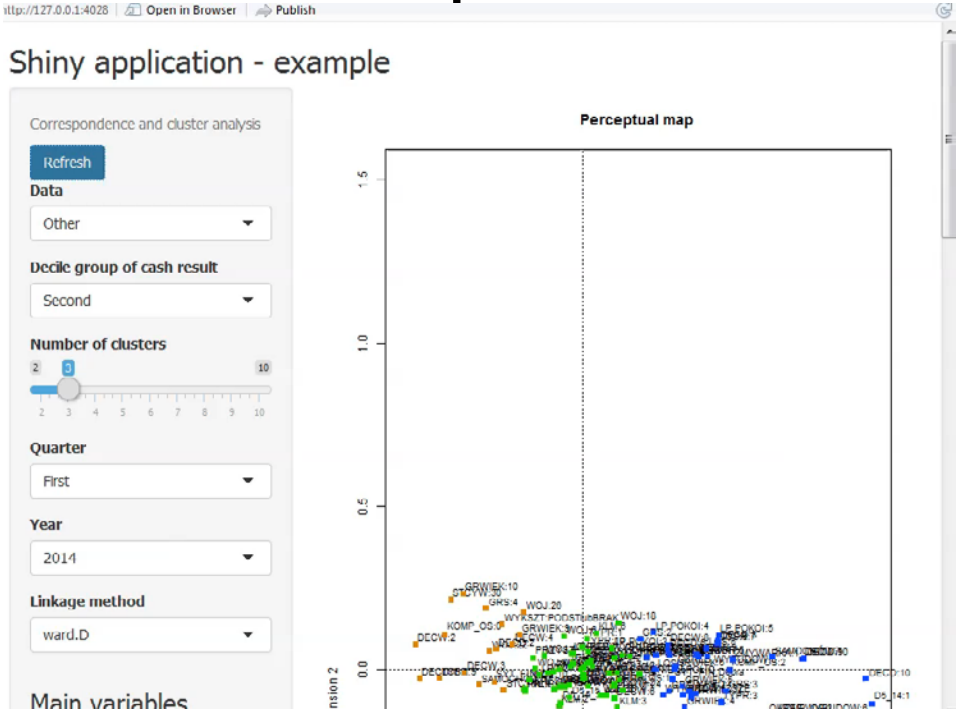
- The dendrogram (left) and the perception map (center) profile index (right) with differently colored clusters obtained by average linking method. **Savings** generating households with mortgage in CHF in 2014.



Still no difference…

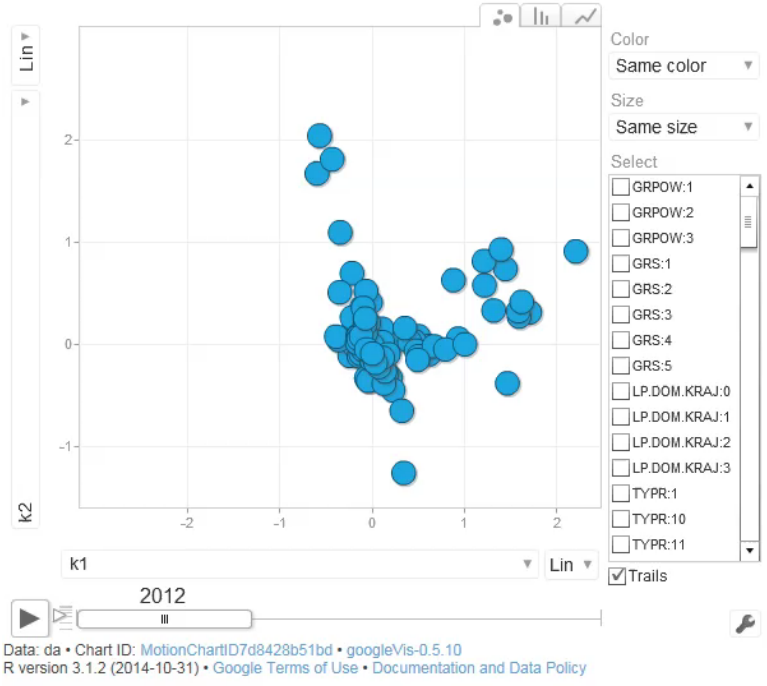
Source: Own calculations.

Clustering for households with mortgages extended for quarterly data with automated presentation in R/Shiny (1 of 2)



Bingo? Seems like extension for the life conditions and quarterly data reveal some seasonal patterns for incomes and expenses.
Extra:
Mortgage bearing households seem to show different clusters.

Clustering for households with mortgages extended for quarterly data with automated presentation in R/Shiny (2 of 2)



More suitable dynamic chart based on GoogleVis for observing e.g. time effects

Conclusions of the results

- The initial aim to find the small sets of (optimally) interpretable features for the most saving and most indebting households was not achieved.
- The results of the HBS based cluster and correspondence analysis applied on 99 transformed features of employees' households show no difference between two marginal types of households in terms of their annual cash results.
- Some seasonal patterns occurred, while the outliers remained in clusters despite further narrowing of the input dataset.
- The differences in clusters between mortgage indebted and not indebted households finally occurred in quarterly data too, however, not yet interpretable for authors at this stage of analysis.
- The results may suggest to extend the number of features rather than searching among the existing 99. (NEWS! from the last moment: Extension for life conditions related features helped!)

Initial evaluation of the statistical tool

- Initial preparation from scratch took around 400 hours of literature review, coding in R and ordered documenting of the knowledge and experience database.
- The tool suits for different input data, e.g. other surveys, administrative data and matched datasets.
- However, the process of normalisation, decision tree for choosing distance measures, selecting clustering methods and evaluation of its appropriateness still requires further automation.
- May serve in cases of complex imputations from e.g. administrative to survey datasets.
- Perhaps, a last resort help in reduction of potential explanatory variables where e.g. number of parameters is close to or exceeds a number of degrees of freedom, and where e.g. Lasso's fail due to type or size of data.

We protect the value of money

www.nbp.pl